Derivative, Gradient, and Lagrange Multipliers

Derivative

Suppose $f : \mathbf{R}^n \to \mathbf{R}^m$ is differentiable. Its *derivative* or *Jacobian* at a point $x \in \mathbf{R}^n$ is denoted $Df(x) \in \mathbf{R}^{m \times n}$, defined as

$$(Df(x))_{ij} = \left. \frac{\partial f_i}{\partial x_j} \right|_x, \quad i = 1, \ldots, m, \quad j = 1, \ldots, n.$$

The first order Taylor expansion of f at (or near) x is given by

$$\hat{f}(y) = f(x) + Df(x)(y - x).$$

When $y - x$ is small, $f(y) - \hat{f}(y)$ is very small. This is called the *linearization* of f at (or near) x.

As an example, consider $n = 3$, $m = 2$, with

$$f(x) = \left[\begin{array}{c} x_1 - x_2^2 \\ x_1 x_3 \end{array} \right].$$

Its derivative at the point x is

$$Df(x) = \left[\begin{array}{ccc} 1 & -2x_2 & 0 \\ x_3 & 0 & x_1 \end{array} \right],$$

and its first order Taylor expansion near $x = (1, 0, -1)$ is given by

$$\hat{f}(y) = \left[\begin{array}{c} 1 \\ -1 \end{array} \right] + \left[\begin{array}{ccc} 1 & 0 & 0 \\ -1 & 0 & 1 \end{array} \right] \left(y - \left[\begin{array}{c} 1 \\ 0 \\ -1 \end{array} \right] \right).$$

Gradient

For $f : \mathbf{R}^n \to \mathbf{R}$, the *gradient* at $x \in \mathbf{R}^n$ is denoted $\nabla f(x) \in \mathbf{R}^n$, and it is defined as $\nabla f(x) = Df(x)^T$, the transpose of the derivative. In terms of partial derivatives, we have

$$\nabla f(x)_i = \left. \frac{\partial f}{\partial x_i} \right|_x, \quad i = 1, \ldots, n.$$

The first order Taylor expansion of f at x is given by

$$\hat{f}(x) = f(x) + \nabla f(x)^T (y - x).$$

Gradient of affine and quadratic functions

You can check the formulas below by working out the partial derivatives.

For f affine, *i.e.*, $f(x) = a^T x + b$, we have $\nabla f(x) = a$ (independent of x).

For f a quadratic form, *i.e.*, $f(x) = x^T P x$ with $P \in \mathbf{R}^{n\times n}$, we have $\nabla f(x) = (P + P^T)x$. When P is symmetric, this simplifies to $\nabla f(x) = 2Px$.

We can use these basic facts and some simple calculus rules, such as linearity of gradient operator (the gradient of a sum is the sum of the gradients, and the gradient of a scaled function is the scaled gradient) to find the gradient of more complex functions. For example, let's compute the gradient of

$$f(x) = (1/2)\|Ax - b\|^2 + c^T x,$$

with $A \in \mathbf{R}^{m\times n}$. We expand the first term to get

$$f(x) = (1/2)x^T(A^T A)x - b^T Ax + (1/2)b^T b + c^T x,$$

and now use the rules above to get

$$\nabla f(x) = A^T Ax - A^T b + c = A^T(Ax - b) + c.$$

Minimizing a function

Suppose $f : \mathbf{R}^n \to \mathbf{R}$, and we want to choose x so as to minimize $f(x)$. Assuming f is differentiable, any optimal x (and it's possible that there isn't an optimal x) must satisfy $\nabla f(x) = 0$. The converse is false: $\nabla f(x) = 0$ does not mean that x minimizes f. Such a point is actually a *stationary point*, and could be a saddle point or a maximum of f, or a local minimum. We refer to $\nabla f(x) = 0$ as an *optimality condition* for minimizing f. It is necessary, but not sufficient, for x to minimize f.

We use this result as follows. To minimize f, we find all points that satisfy $\nabla f(x) = 0$. If there is a point that minimizes f, it must be one of these.

Example: Least-squares. Suppose we want to choose $x \in \mathbf{R}^n$ to minimize $\|Ax - b\|$, where $A \in \mathbf{R}^{m\times n}$ is skinny and full rank. This is the same as minimizing $f(x) = (1/2)\|Ax - b\|^2$. The optimality condition is

$$\nabla f(x) = A^T Ax - A^T b = 0.$$

Only one value of x satisfies this equation: $x_{\text{ls}} = (A^T A)^{-1} A^T b$.

We have to use other methods to determine that f is actually minimized (and not, say, maximized) by x_{ls}. Here is one method. For any z, we have

$$(Az)^T(Ax_{\text{ls}} - b) = z^T(A^T Ax_{\text{ls}} - A^T b) = 0,$$

so $Az \perp Ax_{\rm ls} - b$. Now we note that

$$\begin{array}{rcl} \|Ax-b\|^2 & = & \|Ax_{\rm ls} - b + A(x-x_{\rm ls})\|^2 \\ & = & \|Ax_{\rm ls} - b\|^2 + 2(A(x-x_{\rm ls}))^T(Ax_{\rm ls}-b) + \|A(x-x_{\rm ls})\|^2 \\ & = & \|Ax_{\rm ls} - b\|^2 + \|A(x-x_{\rm ls})\|^2 \\ & \geq & \|Ax_{\rm ls} - b\|^2 \end{array}$$

using the orthogonality result above. So this shows that $x_{\rm ls}$ really does minimize f. With this argument, we really didn't need the optimality condition. But the optimality condition gave us a quick way to find the answer, if not verify it.

Lagrange multipliers

Suppose we want to solve the constrained optimization problem

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & g(x) = 0, \end{array}$$

where $f : \mathbf{R}^n \to \mathbf{R}$ and $g : \mathbf{R}^n \to \mathbf{R}^p$.

Lagrange introduced an extension of the optimality condition above for problems with constraints. We first form the *Lagrangian*

$$L(x,\lambda) = f(x) + \lambda^T g(x),$$

where $\lambda \in \mathbf{R}^p$ is called the *Lagrange multiplier*. The (necessary, but not sufficient) optimality conditions are

$$\nabla_x L(x,\lambda) = 0, \qquad \nabla_\lambda L(x,\lambda) = g(x) = 0.$$

These two conditions are called the KKT (Kharush-Kuhn-Tucker) equations. The second condition is not very interesting; we already knew that the optimal x must satisfy $g(x) = 0$. The first is interesting, however.

To solve the constrained problem, we attempt to solve the KKT equations. The optimal point (if one exists) must satisfy the KKT equations.

Example: Linearly constrained least-squares. Consider the linearly constrained least-squares problem (see lecture slides 8)

$$\begin{array}{ll} \text{minimize} & (1/2)\|Ax-b\|^2 \\ \text{subject to} & Cx - d = 0 \end{array}$$

with $A \in \mathbf{R}^{m\times n}$ and $C \in \mathbf{R}^{p\times n}$. The Lagrangian is

$$\begin{array}{rcl} L(x,\lambda) & = & (1/2)\|Ax-b\|^2 + \lambda^T(Cx-d) \\ & = & (1/2)x^T AAx - b^T Ax + (1/2)b^T b + (C^T\lambda)^T x - \lambda^T d. \end{array}$$

The KKT conditions are

$$\nabla_x L(x, \lambda) = A^T Ax - A^T b + C^T \lambda = 0, \qquad \nabla_\lambda L(x, \lambda) = Cx - d = 0.$$

These are a set of $n + p$ linear equations in $n + p$ variables, which we can write as

$$\begin{bmatrix} A^T A & C^T \\ C & 0 \end{bmatrix} \begin{bmatrix} x \\ \lambda \end{bmatrix} = \begin{bmatrix} A^T b \\ d \end{bmatrix}.$$

If the matrix on the left is invertible, this has one solution,

$$\begin{bmatrix} x^\star \\ \lambda^\star \end{bmatrix} = \begin{bmatrix} A^T A & C^T \\ C & 0 \end{bmatrix}^{-1} \begin{bmatrix} A^T b \\ d \end{bmatrix}.$$

As in the least-squares example above, you have to use another argument to show that $x^\star$ found this way actually minimizes f subject to $Cx = d$. We don't expect you to be able to come up with this argument, but here's how it goes. Suppose that z satisfies $Cz = 0$. Then

$$(Az)^T (Ax^\star - b) = z^T (A^T Ax^\star - A^T b) = z^T (-C^T \lambda^\star) = -(Cz)^T \lambda^\star = 0,$$

so $(Az) \perp (Ax^\star - b)$. Using exactly the same calculation as for least-squares above, we get

$$\|Ax - b\|^2 \geq \|Ax^\star - b\|^2,$$

which shows that $x^\star$ does indeed minimize $\|Ax - b\|$ subject to $Cx = d$.